OVERVIEW: INFORMATION EXTRACTION FROM BROADCAST NEWS

Patricia Robinson (1), Erica Brown (2), John Burger (1), Nancy Chinchor (2),
Aaron Douthat (2), Lisa Ferro(1), Lynette Hirschman (1)

(1) The MITRE Corporation
202 Burlington Rd.
Bedford, Massachusetts 01730
(2) Science Applications International Corporation
10260 Campus Pt. Dr. M/S A2-F
San Diego, California 92121

ABSTRACT

Broadcast news is a rich domain for information extraction, but one that presents new challenges for evaluation. In this paper we present an overview of the first evaluation of information extraction from broadcast news that was conducted as part of the DARPA-funded Hub 4 1998 workshop. We discuss the work that was required to design and administer the evaluation, describe some of the challenges that we encountered, and summarize the results of the evaluation.

1. INTRODUCTION

For over a decade the Tipster Program has sponsored the Message Understanding Conferences to evaluate information extraction in the newswire domain. The Hub 4 and Hub 5 workshops have also received considerable attention for their work in the evaluation of automatic speech recognition technology. This year's DARPA Broadcast News evaluation pioneered the formal evaluation of information content for continuous speech recognition systems. For the first time, systems were evaluated in terms of an information extraction metric in addition to the conventional Word Error Rate (WER). This additional evaluation offered new perspectives on the problems of continuous speech recognition, and the ways in which information extraction interacts with it.

FOR NATIONAL PUBLIC RADIO THIS IS JOE SMITH IN CLEVELAND A DELTA AIRLINES JETLINER SLIT OFFER RUNWAY CLEVELAND SNOWY HOPKINS INTERNATIONAL AIRPORT LAST NIGHT NO ONE WAS INJURED HE WAS THE SECOND SUCH INCIDENT IN AS MANY DAYS

Figure 1: Noisy speech transcription

2. THE NAMED ENTITY TASK

The goal of the Named Entity task is to identify named expressions in broadcast news transcriptions. Because the transcriptions are automatically generated by ASR systems, they often contain errors of insertion, deletion, or substitution. The transcript in Figure 1, for example, is noisy (speech recognition errors are underlined). The ASR system erroneously recognized "*slid off a runway at cleveland's*" as "*slit offer runway cleveland.*" However, interesting information, e.g., named entities, can still be extracted from such transcriptions.

Systems identify named entities by inserting SGML annotations into the speech transcript. These annotations are divided into three categories: ENAMEX (Entity Name Expressions), NUMEX (Number Expressions), and TIMEX (Time Expressions). ENAMEX consists of person, organization, and location names. NUMEX and TIMEX designate money and percent expressions,

FOR <enamex type=org>NATIONAL PUBLIC RADIO</enamex> THIS IS <enamex type=per>JOE SMITH</enamex> IN <enamex type=loc>CLEVELAND</enamex
A <enamex type=org>DELTA AIRLINES</enamex> JETLINER SLIT OFFER RUNWAY <enamex type=org>CLEVELAND SNOWY HOPKINS</enamex> <enamex type=loc>INTERNATIONAL AIRPORT</enamex> LAST NIGHT NO ONE WAS INJURED HE WAS THE SECOND SUCH INCIDENT IN AS MANY DAYS

Figure 2: Noisy speech transcription with Named Entities

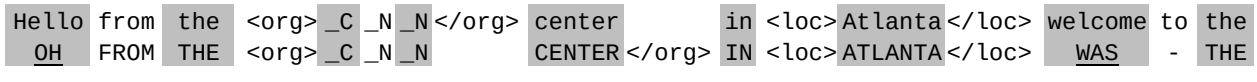

Figure 3: Annotated reference and hypothesis, phonetically aligned
(non-standard SGML is used for brevity, e.g., <org> for <enamex type=org>

and date and time expressions, respectively. Figure 2 shows the same example, annotated by an information extraction system, with namex entity errors underlined.

2.1. Background

In 1995 the Sixth Message Understanding Conference (MUC-6) introduced Named Entity as a component information extraction task. The new task addressed the need for domain-independent extraction of practical value[6]. The Named Entity task was considered to be so successful that the Multilingual Entity Task (MET) was launched to evaluate information extraction on foreign language texts. The first MET (MET-1) focused on Spanish, Chinese, and Japanese[4]. The second MET (MET-2) evaluated Named Entity extraction from the Chinese and Japanese languages.

In 1998 the Seventh Message Understanding Conference (MUC-7) declared Named Entity extraction from English language newswire articles a well-understood problem. Performance on the Named Entity task is determined via F-measure, which is the combination of precision and recall. The best automatic system performance was measured as approximately 93 F, which compared to human performance of about 97 F.

3. TRANSITIONING TO THE BROADCAST NEWS DOMAIN

In order to transition the MUC-style evaluation to the new domain of broadcast news, a number of adaptations had to be made. This section describes the development of a new scorer to address noisy speech input, the definition of new task guidelines for Hub 4, and the data annotation process.

3.1. Alignment and Scoring

The scoring of an extraction system response against a reference transcript marked with "ground truth" named entities presents a number set of problems. The scoring procedure must allow for differences not only in named entity annotation but also in the underlying word stream. As shown above, system-produced transcription is adversely affected by such factors as noise in the production, transmission and capture of the audio signal and an errorful transcription process.

MITRE's mscore program[1] made use of a phonetic alignment procedure[3] to mediate between system transcription and reference transcription. SAIC incorporated this algorithm with the MUC Named Entity algorithm to produce the Hub 4 Named Entity scoring algorithm that was used in this year's Broadcast News workshop. The alignment step is crucial in this process, because of the differences in the underlying text stream just discussed. Figure 3 shows a portion of an annotated reference transcription, aligned phonetically with a system's output. The transcription errors made by the ASR system are underlined. Word-level alignment enables the NE scoring algorithm to determine which reference named entities and hypothesis (system) named entities to compare for scoring purposes.

The scoring algorithm compares the extracted information in the aligned data in three dimensions: extent (locating the correct region), type (assigning the right class to the extracted elements) and content (getting the underlying words right). For audio data, it is important to separate a content score (closely related to word error) from scores for extent and type. This permits a system to get credit for identifying a region in which there is something interesting (e.g., a name), even where the speech recognizer mis-transcribes the region. In the example, we can see that the extent of the named entity _C_N_N CENTER is incorrect, but its type, an ORG, is correctly annotated.

Separating the type, extent and content measures supports research in audio indexing and browsing, i.e., can the system identify key regions to listen to, as well as research on the use of prosodic information independent of transcription accuracy.

3.2. Task Definition

In order to maintain comparability across evaluations and preserve historical information extraction conventions, MITRE and SAIC changed the MUC-7 Named Entity task definition only minimally for Hub 4[2]. However, the guidelines were refined to increase interannotator agreement in some areas. For example, the consistent annotation of complex relative temporal expressions, such as "*since the november crash*" proved to be difficult for MUC-7 annotators. As a result, Hub 4 restricted the annotation to more straightforward absolute temporal

expressions. In addition, the task definition and annotation guidelines required revision to account for the language phenomena and format particular to the broadcast news domain, as described in the following sections.

Disfluencies.

Disfluencies, such as word fragments, repetitions, restarts, and pause fillers, are prevalent in the broadcast news domain, and even more so in conversational speech. As a result, content annotators were forced to develop annotation guidelines for named entities that contained disfluencies. It was decided, for example, that in the case of a sequence of repetition or restarts, that the last full mention of a named entity would get marked: "*I only go go to lebanon into lebanon lebanon lebanon* <location>*lebanon*</location>".

Lack of capitalization.

Lack of capitalization information in speech transcripts complicates the task of determining phrase extent for both human annotators and systems. Without world knowledge, for example, it is difficult to determine the extent of the location name "*newport beach."* The task definition encouraged annotators and system developers to consult *Merriam Webster's Geographical Dictionary* to disambiguate phrase extent for location names[5].

Tokenization.

The two evaluation areas in this year's Hub 4, recognition accuracy and information extraction, place very different demands upon the data preparation process. It became apparent during the design of the evaluation, and the subsequent preparation of data, that these demands were sometimes in sharp conflict. The most salient example was the word-level segmentation conventions required for the calculation of WER that conflicted with the tokenization conventions inherent to information extraction markup. For instance, for purposes of WER, it is perfectly sensible that the genitive ending (apostrophe-s) is not a separate token, but merely part of another word. Thus, "*bill clinton's recent visit*" comprises four words. However, the convention in information extraction annotation has been that the main word can be separated from such an ending, e.g., "<person>*bill clinton*</person>*'s recent visit*", effectively forming five tokens. One manifestation of this problem is that it is impossible to encode these differing segmentations with SGML, due to crossing extents. A solution of sorts (the so-called pseudo empty tags) addressed this issue, but it in turn made it difficult to automatically validate the annotation.[1]

3.3. Data Preparation

Corpus Annotation.

Annotators at MITRE and SAIC tagged approximately 40 hours of broadcast news transcripts which were produced by the Linguistic Data Consortium. Annotators used a version of MITRE's Alembic Workbench (AWB) annotation tool that was adapted to support the new SGML format required for the evaluation. The AWB tool proved to be critical to consistent, cost-effective annotation. With tools, the Hub 4 annotation costs were tractable; the double annotation and reconciliation of the 40 hour corpus was completed in two staff months.

Annotators at GTE/BBN contributed another 100 hours of annotated data. This data was released to the evaluation participants for training purposes.

Interannotator Agreement.

It is standard practice in the information extraction community to conduct interannotator agreement experiments to verify that a corpus has been annotated consistently and to ensure that the task guidelines are unambiguous. One type of interannotator agreement experiment measures human performance on the Named Entity task by comparing one annotator's tagged document to the final version of document that has been reconciled by at least two annotators. Interannotator agreement for Hub 4 was measured at approximately 98 F. This compares favorably with MUC-7 interannotator agreement, which was measured at about 97 F.

4. EVALUATION RESULTS

The broadcast news domain presents new challenges for information extraction systems. The content and genre of broadcast news makes the extraction task more difficult. For example, speech transcripts lack information that is commonly found in newswire, such as:

---

[1] "Pseudo-empty tags" is a simplistic way to deal with SGML's requirement that elements must nest properly. Under this scheme, start and end tags of named entity elements are both transformed into empty tags, e.g., <b_enamex type=LOC>CLEVELAND<e_enamex>.

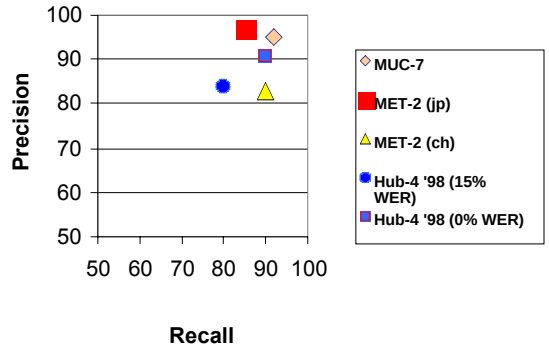

Figure 4: Results of best systems' across evaluations

- Cues, such as titles and corporate designators, which provide contextual evidence to indicate Named Entities.
- Capitalization, which simplifies the task of determining phrase extent.
- Punctuation, which facilitates the task of document zoning, e.g., determining sentence and phrase boundaries.
- Grammaticality, such as well-formed, regularized language, which supports pattern-based information extraction.

However, the results of the Hub 4 evaluation were promising. The chart in Figure 4 summarizes the results of the best systems' Named Entity performance across the MUC-7, MET-2 (Japanese and Chinese), and Hub 4 (extraction on clean transcript and extraction on ASR transcript). We see that with a perfect transcript (0% WER), the best system's named entity scores are only slightly lower (91 F) than for the best system's named entity scores on MUC-7 newswire (93 F). With increased word error rate, named entity performance degrades. However, at 15 percent WER, the best system's named entity performance is still a respectable 82 F.

5. CONCLUSION

The first evaluation of information extraction from broadcast news was successful. The performance of the best information extraction systems suggests that useful extraction can be performed even from noisy ASR transcripts. Infrastructure, such as the scoring pipeline and annotation tools, are in place to support future evaluations.

REFERENCES

1. Burger, J., Palmer, D., and Hirschman, L. "Named Entity Scoring for Speech Input," *COLING-98*, Montreal, 1998.

2. Chinchor, N., Brown, Robinson, P. "The HUB-4 Named Entity Task Definition (version 4.8). Available by ftp from www.nist.gov/speech/hub4_98."

3. Fisher, W. and Fiscus, J. "Better Alignment Procedures for Speech Recognition Evaluation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1993, Vol. II.

4. Merchant, R., Okurowski M. E., Chinchor, N. The Multilingual Task (MET) Overview. In *Proceedings of the Tipster Program Phase II*, May 6-8, 1996. Tysons Corner, Virginia.

5. *Merriam Webster's Geographical Dictionary Third Edition.* Merriam-Webster, Incorporated. Springfield, Massachusetts.

6. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufman Publishers, San Francisco, CA, November 1995.